T-61.5020 Statistical Natural Language Processing

Exercises 2 — Similarity measures
Version 1.0

1. While waiting for springly sniffles Teemu T. Teekkari tested some medicines. He tried three candidates, which were the Tintus cough syrup, Koskisen Korvalääke and Otaniemen Termiitti. For each of them, he tried to describe how the drug tasted. An official observer wrote down how many times the five chosen adjectives were mentioned for each of the drugs. Results are in the Table 1.

	fresh	acidic	sweet	fruity	soft
Tintus	0	0	5	1	4
Korvalääke	10	6	2	1	0
Termiitti	1	4	3	3	3

Table 1: Document-word matrix

   Calculate pairwise distances between the drugs using the following measures:

   a) Euclidean distance (L_2 norm)
   b) L_1 norm
   c) Cosine distance
   d) Information radius

   Why Kullback-Leibler divergence is not a practical measure in this case?

2. Let us consider the following measures:

   a) Kullback-Leibler divergence
   b) Information radius
   c) L_1 norm

   If the distance is minimum according to one of the measures, does it mean that it is minimum also according to the others?

3. Consider the distance measures in Problem 2. For each measure, find what kind of distributions would give the largest possible distance. *Hint: For information radius, the largest possible distance is* $2 \log 2$.